Cognitive Neuroscience of Causal Reasoning

Joachim T. Operskalski, Aron K. Barbey

Decision Neuroscience Laboratory, Beckman Institute, University of Illinois, Urbana, IL, USA

Draft to appear in the *Oxford Handbook of Causal Reasoning*

Running Head: Causal Reasoning

Address for correspondence:

Decision Neuroscience Laboratory
Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign
405 North Mathews Avenue
Urbana, IL 61801

Email: Barbey@Illinois.edu
Web: http://DecisionNeuroscienceLab.org/

[Alice] was quite surprised to find that she remained the same size: to be sure, this is generally what happens when one eats cake, but Alice had got so much into the way of expecting nothing but out-of-the-way things to happen, that it seemed quite dull and stupid for life to go on in the common way.

- *Alice's Adventures in Wonderland (Carroll, 1869)*

1. Introduction

Early in Lewis Carroll's first novella following Alice's journey through Wonderland, the titular character found herself musing about the nature of causal relations, trying to explain and predict events based on the curious ties that connect events in a world seemingly unbound by the usual rules of possibility. She correctly attributed her mysteriously shrinking to the fact that she drank a potion marked "drink me," and then correctly predicted that eating cake marked "eat me" might change her size yet again. Even in fiction, when the lines between possible and impossible can be flexed to resemble nothing like those in reality, human thought is still constrained by characteristic patterns of induction and reasoning; anything else ceases to be human thought as it is usually framed in cognitive psychology.

Similarly, the beliefs of real people can be just as far from ground truth as Alice's, but they are still constrained by a set of processes in the mind that are implemented in the brain to support beliefs and goal-directed behavior. Just like Alice – who was surprised by her shrinkage before attributing it to the one novel event immediately preceding it – people typically attribute their physical maladies to novel events and behaviors that break from usual habits. Anecdotes abound of people who can no longer tolerate a specific brand or type of alcohol after a particularly painful hangover; operant conditioning processes may explain the physical aversion, but causal reasoning is required to make sense of it, explaining previous hangovers and taking action to prevent them in the future.

However, causal judgment often requires more than identifying novel events that occur together, especially in the complicated world we inhabit, where multiple variables interact with one another to cause or enable some events while preventing others. Statistical patterns of co-occurrence can be probed to correctly infer causality in many cases. Alice's dramatic growth in Wonderland would have been neither remarkable nor attributable to the magic potion if she had only grown slowly over the course of the next ten years; the base rate of normal growth in children is on the order of several centimeters per year, and would have been expected even without the potion. This is the basis of the probabilistic contrast models of causal judgment: a given relation's causal power is the difference between the probability of seeing an effect in the presence of its purported cause, and the probability of seeing that effect without the cause, separately accounting for the base rates of the events in question (Cheng & Novick, 1990; see Cheng, 1997 for further discussion and criticism of probabilistic contrast in the "Power PC Model").

The aim of this chapter is to discuss causal reasoning from a perspective grounded in neuroscience. Causal reasoning can be studied in the abstract, as a guideline for how to rationally form beliefs and update them, but it can also be studied "in the wild," as it is practiced in the sciences and in daily life, and with attention to the brain and its causality-perceiving mechanisms that result from millions of years of natural selection for more successfully surviving and reproducing systems. Far from only an abstraction of statistics and mathematics, causal reasoning is all around us; it is in essence what field epidemiologists do to explain outbreaks of illness and predict their trajectories in a population; they look at patterns of dependency between events. To explain an individual person's cough, for example, there may be multiple possible causes under consideration: a common cold virus, lung cancer, or heartburn (see (Tenenbaum et

al., 2011) for this example). The probability of having a cough when experiencing each of those conditions is called its likelihood, and the likelihood of a cough is much higher with a virus or lung cancer than it is for heartburn. The probability of each hypothesis being true before seeing any evidence of it is called its prior probability, and the prior for cold viruses and heartburn is much higher than that for lung cancer. Considering all three hypotheses using Bayes' Theorem (using both prior probability and likelihood) then favors the cold virus as the most likely causal explanation for having a cough. Even physicians well-trained in medical diagnostics, however, are prone to ignoring the base rates of rare events when judging whether a positive test result is more likely to be due to disease or an error in testing (Krynski & Tenenbaum, 2007). Although people may possess the raw information processing power to calculate prior and posterior probabilities when instructed how to do so, there is clearly another, more intuitive, set of mental processes available to those who are untrained in statistics and logic. Although they can also lead to errors in belief, intuitive causal judgment processes are far from a flawed way of thinking about the world, especially when patterns of co-occurrence are inadequate to mentally separate causes from their effects. Without complete knowledge of all possible causal factors that could be operating in the background, the probabilistic contrast and other statistical theories of causal judgment are unable to differentiate causes from events that simply co-occur due to a third causal factor. They are also unable to account for the fact that people make causal judgments about events never seen before, and then use those judgments in subsequent reasoning without any possible knowledge of base rates or patterns of co-dependency.

To understand our collective places in the world as both objects and effectors of change, it is necessary to recognize the generative mechanisms linking events that occur in a particular sequence. To then behave in a goal-directed manner (pursuing some ends and avoiding others), it

is also necessary to use and manipulate such knowledge and beliefs about the generative mechanisms that have already been inferred. By mentally representing the world as it is, while also imagining the world as it is not, we are able to integrate new and surprising information with the entirety of our prior experiences to explain the past and predict the future. In other words, we create new knowledge by combining and manipulating prior knowledge. This is the basis of reasoning and judgment.

A remarkable body of work in the cognitive sciences has been devoted to modeling the reasoning process at the behavioral level. Competing models of "rational" or "normative" reasoning describe different ways to make judgments and combine beliefs to generate new ones, and they are typically evaluated for their ability to converge with the solutions generated by the norms of probability theory. Cognitive psychologists have also developed "descriptive" models of reasoning that purport to characterize how lay people actually reason, irrespective of whether that involves converging with theoretical norms.

Strong programs of research are devoted to the study of both sorts of reasoning models in cognitive science and psychology without appealing to the knowledge or assumptions of neuroscience. Rationality and human thought are conceptually self-contained; that is, rationality can be studied using our assumptions on the nature of truth and the rules of formal logic, and human thought can be probed by asking people to solve reasoning problems and asking them what they believe, without ever trying the separate the brain from the behavior or asking how the brain is operating in the background. Beyond demonstrating that the human mind is a manifestation of the structure and function of the brain, then, what could neuroscience possibly contribute to the study of reasoning that isn't equally or more-thoroughly addressed by the experiments and proofs from psychology and philosophy?

More generally, this question (or criticism) could be leveled at the entire field of cognitive neuroscience. To answer it, we also consider three more direct questions:

- What is the goal of cognitive neuroscience?
- How is cognitive neuroscience conducted?
- What can cognitive neuroscience contribute to programs focused purely on cognition or neuroscience alone?

One view of cognitive neuroscience is that it is a merging of already-mature disciplines. By combining principles of behavioral science and neurobiology with norms from probability theory and concepts of truth from philosophy of science, we aim to gain a fuller picture of the nature of truth and the simultaneously powerful and limited way that the human mind understands its environment. The driving goal of cognitive neuroscience is thus to describe how the properties of the brain support the intricate inner workings of the human mind. What makes a human brain different from the simple neural networks of lobsters or sea snails? What makes a modern human brain different from those of modern gorillas and chimpanzees, or the now-extinct homo neanderthalis or erectus? We look at loss-of-function studies and neuroimaging experiments to answer very basic questions about brain-behavior relationships as a whole, and in so doing we gain insight about the component parts as well. Therein lies the possibility of learning how individual neurons represent information in a way that supports representational thought, and why skin cells or muscle cells signaling to one another do not have the same capability. Therein also lies the possibility of learning about the nature of thought itself; by probing the brain's limits of processing power, we learn about the most likely calculation being used in addition to the nature of the cognitive task being engaged: in this case, causal reasoning.

The methods favored by cognitive neuroscientists involve using brain imaging methods to measure the structural and functional correlates of specific psychological events like memory retrieval, and how they explain inter-individual differences in competencies like the number of

items that can be remembered or the ability to inhibit attention to distracting information. Simple statistical tests can be used to show a correspondence between focal brain damage and categorical deficits on very specific information processing abilities. Machine learning algorithms can be used to extract complex patterns of network activity in the brain that correspond to subtle differences in the same abilities. Reviewing the cognitive neuroscience literature on any given topic typically yields a map of brain regions where changes in blood flow, electrical field, or structural integrity correspond to some psychological function of interest. It is tempting, then, to survey the cognitive neuroscience literature on reasoning, combine the results onto a template brain image, and declare that we have uncovered the "reasoning network." Doing so is certainly a promising beginning to our foray into the neuroscience of reasoning, if for no other reason than to make a list of other psychological functions supported by such a network, to then be tested for their possible involvement in the reasoning process as well. However, relying too heavily on a map of task activations from univariate neuroimaging studies will only take us so far in trying to understand the neural mechanisms of reasoning; doing so would ignore the fact that maps of brain activation can be engaged by nuisance variables or "demand characteristics" just as easily as the task of interest, even when the underlying experiments were conducted with rigorous control conditions. The "reverse inference" problem inherent to trying to explain exploratory cognitive neuroscience findings is that a particular brain region or network's ability to support a given cognitive function does not imply that there is only one function served by that region, or that the cognitive function in question is also involved any time its supporting regions are implicated in some other task (Poldrack, 2006). Such an assumption ignores the facts that brain regions support multiple psychological functions, and functionally different brain networks frequently share some nodes in common. Finally, a fundamental functional network to

support some cognitive function of interest will often appear to have moved or changed in its temporal characteristics based on contextual factors other than the function of interest; this insight has led to a theory of intelligence based on a single "multiple demands" network that is the core driver of all facets of goal-directed or intelligent behavior in humans, with differences in brain activations being attributable to demand characteristics, or low level task features like sensory modality or the extent of attention allocation required (Duncan, 2010). With the context and limitations of early cognitive neuroscience methods in mind, we will propose programmatic research on the neural correlates of causal reasoning, with particular attention paid to how we might move beyond univariate task-activation neuroimaging studies.

One motivation for studying reasoning from a cognitive neuroscience perspective could be to engage in the debate between competing models, testing their predictions to offer evidence as to which models are more plausibly being implemented. However, there is a fundamental mismatch between the methods and conceptual canon of the respective fields; the interdisciplinary intersection between the two fields is simply too immature to pursue this end. Cognitive science and neuroscience operate at different levels of conceptual resolution, in that subtle distinctions in symbolic representations of causality have yet to be characterized at a level that can be described in terms of broader patterns of activity in neurons or networks of neurons. Even if the conceptual resolution were made equal, the most basic units of representing information in the study of causality (e.g. truth statements, negation operators) and neuroscience (e.g. action potentials, post-synaptic potentials, blood oxygen level dependent response curves) can not be readily translated into one another. These problems have been referred to as granularity mismatch and ontological incommensurability, as discussed in the context of the mismatch between neuroscience and linguistics (Poeppel & Embick, 2004). Whereas Poeppel

and colleagues suggest using behavior (language, in their case) as a model system to understand computation in the brain, we apply their sentiment to causal reasoning: understanding the form of reasoning sheds light on how the brain represents information. We also believe, however, that understanding the brain's mechanisms can offer insight into the fundamental nature of reasoning, as long as members of the separate disciplines are made adequately aware of their respective assumptions in trying to map findings from one field onto another.

To more directly answer our final question concerning this interdisciplinary field of study, why is neuroscience evidence important to non-neuroscientists? Information processing systems can be described at three levels that were first proposed for the study of visual perception (Marr, 1982). To fully understand the system, it is advantageous to account for its properties at each level of the hierarchy, and neuroscience offers a complementary perspective to those made available by other disciplines. For any information processing system, the calculation at hand or goal to be achieved is the computational level; recognizing objects or categorizing items to support hunting or gathering behaviors is one type of computation. The set of rules for translating input into output is the algorithmic level, in that it functions as a set of instructions that could be carried out by different people, or with some degrees of freedom. Using checklists of necessary and sufficient features to categorize objects is one algorithm; a more effective algorithm is to appeal to underlying reasons for the features to define a category (Murphy & Medin, 1985). Finally, the way a physical system carries out the calculation using a particular algorithm is the implementation level. Different neurons in visual cortex use changes in the rates of their spiking patterns to signal the receipt of an image corresponding to particular colors or shapes. Object recognition computer programs, on the other hand, can implement similar calculations and algorithms using a physical system involving wires and silicon wafers.

What neuroscience has to contribute to the study of causal reasoning is that it is one of few disciplines poised to support a discussion on the implementation level of human reasoning. Those in cognitive psychology who are interested in the descriptive validity of reasoning models clearly need to understand the properties and limitations of the system that is being modeled. Even those among us who are only interested in rational models, however, would benefit from comparing them alongside the descriptive models; this is because it could be of interest to know whether (and if so, when) the solutions to reasoning problems that are naturally generated by the brain perform more accurately or efficiently than those using the steps prescribed by rational models. Brains were shaped by the forces of evolution that simply rewarded the solutions for problems of survival and reproduction. This process involved simple adaptations, like cellular mechanisms for resisting disease, and the behavioral propensity to band together in social groups for support. It also involved the ability to not only learn which parts of the environment were safe or dangerous, but also to predict whether that might change in the future on the basis of new information. This is at the heart of causal reasoning, and we aim to understand how human biology generated a solution to the need for explanation and prediction.

The goals of this chapter are twofold: to survey the current states of the cognitive and neural literatures on causality while acknowledging the mismatch between methods and observations between the disciplines, and to make the case for further developments in the cognitive neuroscience approach to study reasoning, in pursuit of a program from which scholars residing in pure cognitive science and pure neuroscience will benefit equally as those who operate in the intersection between the fields. Toward this end, we will use the following structure:

- Examine the descriptive instantiations of several rational models of reasoning, considering the predictions they might make if they were to be implemented by a neural system as currently understood.
- Review the results of neuroscience experiments aimed at probing how the brain supports the concept of causality, considering whether they have any implications for the differences between purely cognitive models of causal reasoning.

The first model we will consider is Mental Models Theory (MM), which suggests that abstract representations of states of affairs in the world are constructed on the basis of possible co-occurrences of events that are licensed under a particular relation like "cause" or "prevent" (see Johnson-Laird and Khemlani, this volume). Mental Models feature deductive reasoning over fundamentally deterministic relations as the primary method of combining knowledge about separate relations to draw conclusions or generate new knowledge. The second model is Causal Models Theory (CM), suggesting that causal reasoning is supported by abstract representations linking events to one another as a probabilistic network that can be depicted visually with directed graphs and structural equations (see Rottman, this volume). CM theory features the representation of probabilities and inductive reasoning as a central element of causal reasoning. The third model considered in this chapter is Force Composition Theory (FC), in which causal relations are represented in terms of forces interacting with one another to account for the movement of a system toward or away from a particular endstate (see Wolff, this volume). FC theory emphasizes perceptual representations of forces that preserve the structure of the relations being symbolized. Diagrams depicting force vectors are thus used to describe the way individual force representations can be combined to draw conclusions from previously unconnected relations. Causal representations in FC theory depend on an understanding of the way physical forces interact with one another, but are also flexible enough to be analogously applied to more abstract forces like emotion and interpersonal communication.

Many of the behavioral predictions of each theory are identical; they converge on the same inference being drawn in a particular context, which is part of the reason for the granularity mismatch between the psychology and neuroscience approaches to modeling causal reasoning. Cognitive scientists and psychologists draw very fine distinctions between modes of thought, while cognitive neuroscientists are still building theories that map coarse concepts like causal attribution onto large-scale brain networks. The cognitive theories' departures from one another are in the predictions of how people draw inferences in complicated or ambiguous scenarios, so we will focus our discussion on how people combine multiple causal relations to draw inferences about transitivity (or lack thereof). Causal reasoning itself is complex, and has presumably evolved to support representations roughly resembling truth to be drawn from the complex and often inconsistent information about dependencies between events around us. The ability to draw conclusions that only approximate ground truth may be adequate to learn enough about our environment to survive and reproduce, which may account for why the descriptive computational theories of reasoning depart from the rational solution to a reasoning problem at times. A great deal of research in cognitive neuroscience is undertaken with the short-term goal of localizing behaviors to modules or networks in the brain, identifying the common and distinct neural correlates of dissociable psychological functions. A loftier goal of much of the same research (and perhaps a more nebulous one) is identifying the underlying organizational principles that dictate how the physical properties of the brain support the cognitive architecture supporting the mind. From this perspective, understanding the physical implementation of causal reasoning in the brain can help constrain psychological theories to reflect the properties of the biological system supporting it (Goel, 2005). We acknowledge from the outset that the cognitive neuroscience evidence on causal reasoning, rather like the evidence from the behavioral and

cognitive realms, is not free of ambiguity. Further programmatic research, making use of recent developments in neuroimaging technology, holds promise for resolving some of the ambiguity concerning mental constructs that may fundamentally feature a number of alternative modes of thought available under different circumstances.

At the most general level, cognitive neuroscience evidence supports a distributed information processing system engaged by goal-directed behavior, including such networks as a fronto-parietal tract supporting attention and executive control, and fronto-hippocampal tracts supporting memory encoding and retrieval (Barbey, Colom, et al., 2012; Duncan & Owen, 2000; Vincent et al., 2008). Causal reasoning is likely to engage a subset of those networks in the service of goal-directed behavior, and can be subdivided into such processes as judgment or recognition of causality, prediction, and explanation. The different psychological theories of causal reasoning each make predictions about the information processing steps that people use when reasoning over complex sets of relations. Many of those predictions are invisible to neuroscience methods at their current conceptual (and temporal/spatial) resolution, but some of them have implications with respect to the likely neural correlates of reasoning behavior, selectively highlighting elements of the attention, memory, and control networks mentioned above. Furthermore, the nature of the causal representations themselves will also be reflected in the neural correlates of causal reasoning. Currently, it remains unclear whether causal beliefs are supported by simulation mechanisms in the brain that are specific to sensory modalities, abstract semantic knowledge networks, or some combination of the two.

2. Psychological Theories of Causal Reasoning

Statistical patterns can be used to induce a causal relation between events not previously thought to be linked, but this does not account for all instances of causal judgment. It is simple enough to agree, in the context of several decades' worth of medical research, that "smoking causes cancer." Describing the precise nature of the relation as it plays out in specific cases is complicated, however, when such preventing and aggravating factors as poverty, education, diet, stress and genetic inheritance also appear to coincide with both smoking and cancer. In moving from judgment to reasoning (especially under conditions of uncertainty about the original beliefs from judgment processes), the complexity of calculating statistical dependencies rapidly increases with the number of relations being linked as people consider a chain of possibly-linked events. Statistical co-dependency calculations thus account for even fewer instances of reasoning than they do for cases of pure judgment. Far from trying to describe the ideal way to reason about causal relations, descriptive theories of naïve causality emerged from a desire to instead describe actual causal reasoning in daily life.

For drawing conclusions on the basis of some accepted set of premises (the premises being previously accepted causal relations, or a background of prior knowledge), we thus have a family of theories of reasoning that each propose special frameworks for creating causal representations that can be used to map evidence (in the form of events and their co-occurrences) onto novel conclusions.

Before delving into the features of each psychological theory, note the absence of several influential accounts of causal judgment and reasoning in our discussion (see (Ahn et al., 1995; Cheng, 1997; Tenenbaum & Griffiths, 2003) for dependency-based accounts of causal induction and belief updating). Here, we focus on the computational theories that make behavioral predictions about how people represent and combine multiple causal relations to draw

conclusions, especially those for which a plausible neural implementation is available based on current theory of brain function in cognitive neuroscience.

2a. Mental Models Theory

Mental Models Theory (MM) as a representational account of causal reasoning was inspired by the fact that the rules and representational scheme of formal logic result in a combinatorial explosion of additional clauses and statements that must be represented when trying to draw conclusions from some set of events or state of affairs in the world. The key example offered by its original proponents is that claims of the sort "one of these statements is true and the other is false" are much more complex when represented as a series of Boolean algebraic statements than they appear to be when non-logicians think about them. See (Johnson-Laird, 2010a) for the full line of reasoning. The alternative to Boolean algebra and formal logic is a more intuitive solution, and one that resonates with common experience: people are able to compare current states of affairs with possible states of affairs that do not currently exist, while also deciding what sorts of affairs are not possible in the context of what is already believed to exist.

MM theory suggests that people reason by constructing abstract mental representations that license possible co-occurrences of events or states of affairs under a given relation, like "cause," "enable," or "prevent" (Goldvarg & Johnson-Laird, 2001; see Cheng & Novick, 1991 for another model of the difference between cause and enable). The models are modal in the linguistic or psychological sense; the conditions specified in a particular model represent necessary (obligatory) or possible (permitted) states of affairs. Table 1 demonstrates an example of how "cause," "enable," and "prevent" relations between two events or states A and B can be

depicted using models. The capital letters represent a variable, and lowercase letters are used to represent the presence of the variable or event in question. A negation operator (¬a or ¬b) is used to represent the absence of the event in question.

If we consider the relation between smoking and cancer, we can generate mental models that represent possible states of affairs under each type of relation. "Smoking causes cancer" is a general causal statement that licenses three specific possibilities: that smoking and cancer both occurred, that no smoking occurred but cancer occurred anyway due to another cause, and neither smoking nor cancer occurred. And what about the possibility that a person smoked but did not suffer from cancer? It is easy to imagine long-term smokers who never succumbed to the graver consequences typically attributed to cancer. When people are asked to explain why not, they cite preventing factors like protective genetic mutations, rather than claiming that causality is inherently probabilistic. This distinction, that the meanings of causal concepts are deterministic, is a fundamental principle of MM theory. Some prior accounts of causal reasoning do not make strong claims discriminating between the meanings of verbs such as "cause" and "enable" (Cheng & Novick, 1990); they both increase the probability of an effect, and are only used differently based on notions of agency versus circumstance, or background conditions versus a manipulated factor. MM theory suggests that "enable" has a fundamentally different meaning, such that "smoking enables cancer" includes the possibility that smoking occurred but cancer did not, instead of the possibility that cancer occurred independently of cancer. The other two individual models in the set are identical between the two concepts. "Cause" thus refers to conditions of sufficiency to bring about an effect, whereas "Enable" refers to a necessary condition that is not sufficient to bring about an effect on its own. "Prevent," on the other hand, refers to mutually exclusive conditions. "Antioxidants prevent cancer," for example, allows the

possibility of antioxidants being present and cancer being absent, antioxidants being absent and cancer being present, and neither antioxidants nor cancer being present. Multiple causal relations are then combined with one another by listing all of the models (possible co-occurring states of affairs) under each relation in a single set, omitting redundant models, and then removing the middle event or state to link the first and last events.

For each relation in MM theory, the entire set of models allowed is collectively referred to as the fully explicit model for that term. In special cases, concepts like "cause" and "prevent" can take on the strong form of both necessity and sufficiency, such that only the first and the last models listed in each column of Table 1 are considered part of the meaning. If, for example, smoking were the only possible mechanism of developing cancer, then "smoking causes cancer" would take on the strong form of "Cause." Consider substance abuse for a more realistic example: "substance use causes intoxication." There are no other possible causes for intoxication, and substance use (voluntary or involuntary) must occur before intoxication (example taken from (Goldvarg & Johnson-Laird, 2001)).

According to MM theory, people typically construct implicit mental models that capture only a subset of the fully explicit representations. Causal terms are usually used with one implied meaning, so conventions in communication and notions of shared knowledge lead people to assume that the first modality of each set in Table 1 – the implicit model – is the one intended when talking about causal relations. The selection of implicit models is further supported by the *principle of truth*, according to which people naturally represent what is true about a given state of affairs rather than considering the world otherwise. This bias toward representing true states of affairs is the basis for the prediction that combining multiple causal relations will be easier for relations that can be accurately combined while using only the implicit models; those requiring

the fully explicit models and selective removal of prohibited models will be more difficult and prone to error. For example, consider the combination of two "cause" relations, as opposed to two "prevent" relations. Smoking causes lung cancer, and lung cancer causes respiratory problems. The implicit model for both relations contain all of the information needed to combine them, leading to the transitive inference that smoking causes respiratory problems. Double-prevention is more difficult. Healthy habits prevent lung cancer, and lung cancer prevents good health. If we hastily limit ourselves to the implicit mental models, it is tempting to simply drop the middle term – lung cancer – and draw a similar transitive inference that healthy habits prevent good health. Healthy habits clearly do not prevent good health, and not only because the conclusion is inconsistent with experience. "Prevent" relations are not transitive because the second instance of prevention requires the presence of an event that is absent after realizing the first prevention. The erroneous bias toward inferring transitivity has been observed in behavioral experiments among undergraduate students, confirming the prediction of MM theory that double prevention will result in an erroneously transitive conclusion (Goldvarg & Johnson-Laird, 2001). See Table 2 for an example of causal relations (A causes B, and B prevents C) that can be combined using mental models alone, and Table 3 for causal relations requiring fully explicit models to arrive at the correct solution (A prevents B, and B causes C).

In summary, MM theory proposes that causal reasoning depends on multiple deterministic relations – implicit mental models – that can interact to assist or prevent one another from having a particular effect. Therefore, conclusions are drawn in MM by deducing the possible states of affairs that are entailed in the combinations of such deterministic relations in the context of background knowledge.

A Causes B	A Enables B	A Prevents B
a b	a b	a ¬b
¬a b	a ¬b	¬a b
¬a ¬b	¬a ¬b	¬a ¬b

Table 1. Models representing "cause," "enable," and "prevent" relations in Mental Models Theory. Letters represent the presence of an event or state characteristic of the category being represented. The negation operator ¬ represents the absence of the specified event or state, and should be read as "not-a" or "not-b". Each column represents the fully explicit models for each relation, and the first cell in each column represents the corresponding mental model.

A Causes B	B Prevents C	A Prevents C
a b	b ¬c	a b ¬c
¬a b	¬b c	¬a b ¬c
¬a ¬b	¬b ¬c	¬a ¬b c
		¬a ¬b ¬c

Table 2. Fully explicit models representing the combination of two relations – "cause" and "prevent" – to yield a "prevent" relation between the first and last terms. The implicit mental models in the first row are adequate to draw the rational conclusion that A prevents C. Note that dropping the middle terms of the fully explicit models in the third column yields an identical set of fully explicit models linking A and C to those that link any other two prevent relations.

A Prevents B	B Causes C	A Does Not Prevent C
a ¬b	b c	a ¬b ¬c
¬a b	¬b c	a ¬b c
¬a ¬b	¬b ¬c	¬a b c
		¬a ¬b c
		¬a ¬b ¬c

Table 3. Fully explicit models representing the combination of two relations – "prevent" and "cause" – to infer that A and C are not causally related. The absence of B does not preclude the possibility of other causes of C being sufficient. This is counterintuitive if assuming the strong version of "B Causes C," in that B is the only possible cause of C, and its prevention also prevents any of its later effects. Using only the implicit mental models in the first row of the first two columns yields this incorrect inference, which is consistent with the *principle of truth* and the prediction that using fully explicit models to reason places a heavier burden on working memory than most people normally use without being instructed to.

2b. Causal Models Theory

Causal Models Theory (CM) proposes that mental representations of causal knowledge reflect probabilistic relationships. Fundamentally deterministic relationships can be represented using causal models as well, albeit in a probabilistic mental representation due to uncertainty concerning hidden variables. CM theory is based on the construction of abstract mental representations and makes use of the Bayesian Belief Network (hereafter: Bayes Net) as a normative approach to causal induction and causal reasoning (Pearl, 2000).

A Bayes Net is a directed, acyclic graph representing events and the relations between them. Events are represented as circles or nodes, and causal links are represented as arrows between the nodes. "Directed" refers to the asymmetry of a causal relation; changing the status of a cause influences the status of an effect, but changing the status of an effect does not influence the status of its antecedents. "Acyclic" only refers to the fact that the networks are not used to describe closed systems or feedback loops. Each Bayes Net is also accompanied by a series of structural equations describing the relations therein. "A causes B" is thus represented with a cause operator ":=" as "B := A"; the positions on either side of the cause operator are not interchangeable. "Prevent" relations can be represented using a tilde "~" instead of the negation operator "¬," such that "A prevents B" is equivalent to "A causes ~B" or "A causes not-B" being represented by "~B := A". "Enable" relations in CM theory imply that the first event is an

enabling factor that allows another causal factor to have its effect. “A causes B when X enables it” is represented by “B := A, X”. See Figure 1 for a Bayes Net representing possible causes, enablers, and preventing factors influencing influenza infection.

CM theory supports predictive and explanatory reasoning by featuring mental intervention as the primary mode of reasoning; variables in the model can be manipulated to take on values that vary from reality or whose status is unknown. The structural equations representing the statistical dependency between states of connected nodes are then used to imagine how intervening on the causal model will have effects that propagate through the network. The construction of complex causal relations is thus supported by intervening on a mental causal model while also combining the structural equations corresponding to each relation in the network.

One respect in which the CM theory differs from the MM theory is that it predicts that double-preventions will result in the correct inference of “cause” or “allow/enable,” instead of incorrectly inferring “prevent.” Although previous experiments confirmed the predictions of MM theory concerning double-prevention, subsequent studies by other experimenters have found evidence supporting this prediction of CM theory (Sloman et al., 2009). The authors acknowledge, however, that subjects’ reasoning in both studies may be subject to the unintended influence of “atmosphere effects”: that a high rate of a particular answer type being correct in an experiment may lead to perseverative answering when subjects encounter more difficult trials. See (Sloman et al., 2009) for a full discussion of more nuanced differences in predictions distinguishing combinations of “cause” and “allow” or “allow” and “prevent”; these distinctions are valuable in distinguishing between theories at the behavioral level, but they are beyond the scope of the current cognitive neuroscience literature on causal reasoning because noninvasive

brain imaging methods have not yet been used to resolve the difference in representing such similar terms as “prevent” and “cause-not” or “allow-not”.

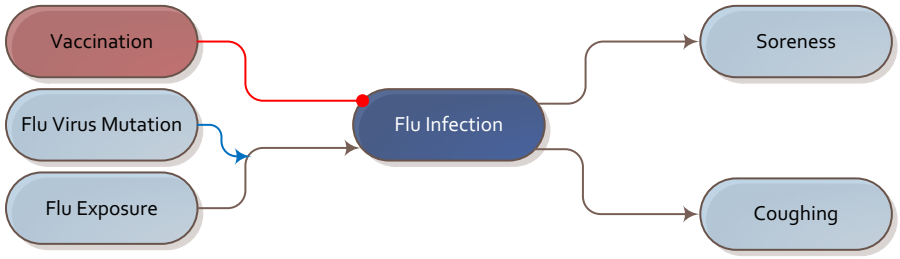


Figure 1. Causal model for influenza
“Cause” relations are represented by grey edges; B := A
“Enable” relations are represented by blue edges; B := A, X
“Prevent” relations are represented by red edges; ~B := A
Flu exposure causes flu infection, unless it is prevented by vaccination. Vaccination can fail, however, if a mutation in the virus enables the exposure to still cause an infection despite vaccination. Note that “enable” relations imply an accessory variable, and do not necessarily over-ride “prevent” relations as depicted in this example.

2c. Force Composition Theory

The third model of causal reasoning we consider is motivated by an effort to ground causal representations in the physical structure of forces and events in the world (Barbey & Wolff, 2002). According to Force Dynamics (or Force Composition, FC) theory, causal relations are mechanistic and represent the transference of a conserved quantity from cause to effect (Ahn & Kalish, 2000). Imagine, for example, a golfer who slices a ball into a tree that knocks it into the hole. Based on understanding the physical mechanisms at work, people would correctly infer that the tree caused the hole-in-one rather than the golfer’s poor shot, even though nobody would claim that trees are generally causes of holes-in-one (Ahn & Kalish, 2000). Whereas prior mechanistic theories did not explain causal relations that take place over a distance (e.g.,

gravity), over large time intervals (e.g., cancer), or with abstract influences that only resemble forces (e.g., social communication), FC theory is flexible enough to account for all such features of causal reasoning because abstract forces can be represented as vectors with magnitude and direction. FC theory is also unique in that its mechanistic representations are concrete: they are grounded in tangible features of the world being simulated. See (Johnson & Ahn, this volume) for current mechanistic theories.

Specifically, causal reasoning in FC theory is supported through the construction of force vectors that represent causal mechanisms between events in the context of tendencies toward or away from an endstate. As with free-body diagrams in Newtonian physics, force vectors are simply iconic arrows with both direction and magnitude that can also be re-conceptualized as the transfer of energy (or causal influence, if you will). Force vector addition along a single axis can thus be used to characterize the overarching structure linking a series of causal relations in a chain of events, which can then be used to predict the future and explain the past by mentally changing the vectors' directions and magnitudes.

The first vector in a force composition diagram is that of the patient: the thing being acted upon. An affector vector then represents the force imparted by the thing acting on the patient. An endstate vector is a positional vector only, representing the state of affairs being caused or prevented. Predicting the future state of the system is achieved by combining the patient vector with the affector vector; if the resultant points in the direction of the patient's endstate, then the affector is said to have caused the endstate. Negative relations like "prevent" can be represented by the removal of a causal force vector, or the addition of a force vector in the opposite direction from that of the endstate; similarly, the removal of a preventing force vector can result in a "cause" relation (Wolff et al., 2010).

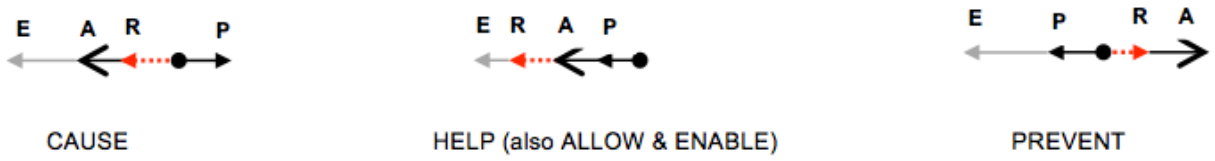


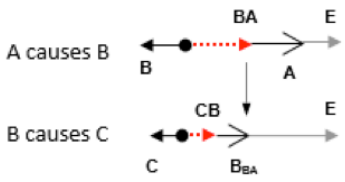
Figure 2. (Reprinted with permission from (Wolff & Barbey, 2015)) Free-body diagrams representing different causal concepts in force composition theory. In these diagrams, **A** = the affector force, **P** = the patient force, **R** = the resultant force, and **E** = endstate vector, which is a position vector, not a force.

Reasoning about complex relations involving multiple patients and affectors is achieved by simply adding vectors from the individual relations. A cause relation is typically represented by a diagram with the patient's vector pointing away from the endstate, the affector's vector pointing toward it, and the difference between the two such that the resultant points toward the endstate. An "enable" or "allow" relation is represented by the patient's vector pointing toward the endstate, the affector's vector pointing in the same direction, with the resultant simply being the superposition of the two. A "prevent" relation involves the patient's vector pointing in a particular endstate's direction and the affector's vector pointing away from it with a magnitude great enough for the resultant to point away from the endstate. Figure 2 depicts the free-body diagrams representing "cause," "enable," and "prevent" relations in FC theory.

When reasoning about two "cause" or "enable" relations, the resultant of the first relation becomes the affector acting on the patient in the second relation. The endstate of the conclusion is taken from the endstate of the second relation. Importantly, the magnitudes of each force vector are relative, rather than explicitly representing a probability. Some combinations of relations are able to support multiple conclusions' being drawn. In those cases, probabilistic relations can be supported by integrating a mathematical function of changes in magnitude on a

particular conclusion being drawn. More specifically, by incrementally changing the magnitudes of the force vectors in each premise relation, the relative frequencies of seeing each conclusion type can be calculated (see (Wolff & Barbey, 2015; supplementary materials) for details on this procedure). For example, combining two prevent relations (A prevents B; B prevents C) will result in a conclusion of either A allows C or A causes C, based entirely on the relative magnitudes of the affector and patient vectors in each premise (see Figure 3). Mentally, the second premise in a double prevention must be represented before the first relation can act on it. Imagine pulling a plug out of a drain in a basin full of water. Pulling the plug out prevents it from being in the drain, and the plug being in the drain prevents the water from leaking out of the basin. Pulling the plug allows/causes the water to leak out of the basin, but mentally representing this set of causal relations requires that there be some concept of water in the basin whose leaking must be allowed in the first place (see (Wolff & Barbey, 2015) for this example and a full discussion of representing double prevention with force vectors).

Note that abstract causes like those involving interpersonal communication (e.g., a compliment causes a student to feel good, and criticism causes a student to feel bad) can also be represented visually using vector arrows that point toward or away from the endstate. The underlying mental representations need not be arrow-based graphs, however, and can still involve the drawing of causal inferences by superimposing such abstract "forces" onto one another.



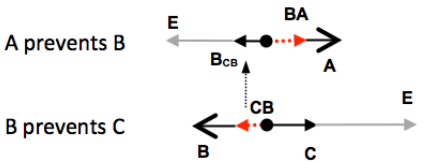


Figure 3. (Reprinted with permission from (Wolff & Barbey, 2015)) Force composition allows the combination of force vectors representing separate causal relations to draw conclusions that are not explicitly represented in the set of causal premises. Combining two CAUSE relations results in the system moving toward the endstate of the second relation. Combining two PREVENT relations similarly results in the system moving toward the endstate in the second relation, with the major difference between the two being the original tendency for the system toward or away from the endstate.

---

3. Neural Implications of Causal Reasoning Models

The reviewed computational models of causal reasoning make alternative claims about the cognitive representations and processes that support causal inference. Each theory further motivates alternative predictions about the neurobiological bases of causal reasoning and can be evaluated in light of the emerging neuroscience literature on causal perception, judgment, and reasoning.

Before turning to a discussion of the different neural implications of each theory, we note commonalities across the reviewed frameworks. Indeed, any model of causal reasoning requires that some information about events and their relations be represented in the mind as a conscious thought. This will require attention mechanisms in the brain to direct one's internal focus toward the piece of information being considered or manipulated; the parietal lobe is typically thought of as a major supporter of attention mechanisms, but modality-specific attention mechanisms are also known to engage perceptual processing substrates in the occipital, temporal and frontal lobes (Smith et al., 2013). The frontal eye field in the posterior frontal lobes, for example, supports eye movement and is involved in directing visual attention; dorsolateral prefrontal cortex (anterior to the frontal eye fields) assists with selective attention, or the ability to attend to some features while suppressing attention to irrelevant ones, along with the many other functions it supports.

Causal reasoning will also require the engagement of memory systems, both to retrieve semantic and episodic memories from previous experience, and to support the ongoing availability of multiple pieces of information during the reasoning process. Memory encoding and retrieval processes are known to rely heavily on subcortical features of the medial temporal lobe, but memory storage and simulation of previous experiences engages the frontal lobes and posterior primary sensory processing networks as well (Buckner et al., 2008).

Lastly, causal reasoning will require the so-called executive control mechanisms in the brain that allow the manipulation of information and selective activation and inhibition of the involved attention and memory processes (Miyake et al., 2000). Executive control processes reliably engage a fronto-parietal network in the brain (Banich, 2009; Vincent et al., 2008).

Taken together, the three broad constructs supporting attention, memory, and executive control processes would suggest that the entire brain is involved in causal reasoning. Where we can separate the component processes from one another, and perhaps separate necessary and sufficient brain networks from those that are simply correlated, is in the predictions made by each of the theories of causal reasoning discussed here, and how they map onto discrete components of the constructs outlined above. The greatest distinction between the descriptive theories of causal reasoning is that they emphasize different underlying modes of information processing. These processing modes can be used to predict the large-scale networks that should be engaged in the brain beyond particular subregions in the prefrontal or medial temporal cortex, which we outline here.

3a. Mental Models Theory

The key to predicting the neural substrates of causal reasoning on the basis of MM theory is the fact that it is based on an abstract, modal code representing possible states of affairs entailed by a causal relation, and uses deductive inference to draw conclusions from multiple relations.

The neural correlates of deductive reasoning have been characterized in prior work as a system of modular brain networks (Goel, 2005). Deductive inference from syllogisms – reasoning from accepted premises to a conclusion guaranteed on the basis of the premises – typically engages a network of frontal and temporal regions. Hemispheric lateralization of deductive reasoning has been observed, but different experimental approaches have resulted in a diverse pattern of findings. For example, studies comparing syllogistic and spatial reasoning against a comprehension control condition revealed a left-sided frontal and temporal reasoning network (Goel et al., 1998). Evidence further indicates that deductive reasoning is localized to the left hemisphere when contrasted against an inductive reasoning task (Goel & Dolan, 2004).

Others studies report different cross-hemispheric dissociations of deductive and inductive reasoning when directly contrasting the two against each other (Osherson et al., 1998). It initially appeared that deductive reasoning engaged a right-sided pattern of parietal activation, which would be consistent with visual and spatial representations being the primary mode of deduction (Osherson et al., 1998). However, when using stimuli that are more easily represented using propositional logic than spatial models (or at least an abstract code), another division of labor emerged; deductive reasoning engaged a right-sided frontal and temporal network (specifically, middle temporal lobe and ventrolateral prefrontal cortex), while inductive reasoning engaged a left-sided fronto-temporal network (specifically, ventrolateral PFC, dorsomedial PFC, insula, posterior cingulate, and the medial temporal lobe) (Parsons & Osherson, 2001).

Some of the differences in results across studies can be attributed to task-specific engagement of brain networks beyond the core reasoning components of a particular task; syllogistic reasoning about unfamiliar information engages perceptual processing regions in the parietal lobe as well, while reasoning about transitive relations concerning familiar semantic content additionally engages medial temporal structures as the hippocampus and parahippocampal gyrus (Goel, 2007; Goel et al., 2000). Further research is necessary to characterize the different contributions of stimulus-specific activation and the true correlates of a core reasoning network, but it is most likely that deduction engages the frontal and parietal cortex, with some right-hemispheric specialization for both spatial and deterministic reasoning.

Proponents of MM theory have described it as an iconic, visuospatial theory of reasoning (Johnson-Laird, 2010b). It employs representations that are iconic in the sense that they preserve the structure or order of events in the world without taking the form of a sensory representation that is isomorphic with the events or objects in the world; events preceding each other in the world are represented by mental models that similarly precede one another in the same order, for example. MM theory is a visuospatial theory of reasoning in that its models can take the form of spatial diagrams. Just as categorical syllogisms can be represented visually with Venn Diagrams or Euler Circles, the possible events that are entailed by a particular causal concept can be represented in the mind using spatial models and set theory notation.

MM theory does not claim, however, to feature mental simulations grounded in the sensory modalities (in this sense, we use "modality" to refer to a particular type of sensory input, rather than its use in modal logic to refer to possibility; any further reference to this meaning will be called "sensory modality" rather than "logical modality"). Whereas deductive inference appears to rely on the left frontal lobes in the brain, the manipulation of spatial objects using

action representations is considered to rely on the parietal lobes (O'Reilly, 2010; Ungerleider & Mishkin, 1982). Human lesion evidence indicates that the right hemisphere is selectively engaged by spatial reasoning, with right posterior cortical lesions (to the parietal, occipital, or posterior temporal lobes) conferring a marked deficit in the mental manipulation of spatial representations, particularly for mental rotation of visual objects (Ratcliff, 1979). Neuropsychological evidence from split-brain patients (whose hemispheres have severely limited communication with one another after surgical resection of the corpus callosum) in addition to neuroimaging evidence measuring the activity in healthy subjects' brains suggest that the right hemisphere is the seat of a visuo-spatial "interpreter" of sorts; both hemispheres are able to perceive visual and spatial information for simple tasks like object identification, but the right hemisphere appears privileged in complex spatial reasoning (Corballis, 2003). Right-hemisphere dominance has thus been suggested as a possible organizing scheme in the brain's implementation of causal reasoning (Johnson-Laird, 1995).

Recent evidence confirms that posterior cortex plays a role in spatial intelligence, but the picture is now more complicated than the right-hemisphere hypothesis suggested previously. Bilateral parietal cortex was engaged by a task requiring spatial discrimination in a neuroimaging study, while mental rotation of spatial objects engaged only the left inferior parietal cortex and right subcortical nuclei (Alivisatos & Petrides, 1997). Another study found bilateral parietal activation during mental rotation of a variety of visual objects (Jordan et al., 2001). The original right-hemisphere reasoning hypothesis (Johnson-Laird, 1995) has thus been updated to suggest instead that the visuospatial processing system in the brain – including primary visual cortex in the occipital lobe and the correlates of higher-order cognitive manipulation in parietal cortex – forms the core of a reasoning network (Goel, 2005). Evaluating this prediction in light of the

neuroscience evidence on reasoning about syllogisms (i.e. explicitly focusing on deductive reasoning) confirms that some types of reasoning engage a visuospatial system, but a number of experimental manipulations also yield engagement of other systems in the brain, suggesting that a dual-process framework of belief activation and evaluation of evidence might be the most accurate way to describe the neural correlates of deduction (Fugelsang & Thompson, 2003).

An intuitive system involving the neural correlates of emotion, language and memory is engaged by the use of heuristics and biases from prior beliefs when reasoning about familiar premises and evidence that is consistent with previous experience, whereas a slower, reflective system involving visuospatial manipulation is engaged by reasoning about unfamiliar premises or those involving a conflict between evidence and belief (Goel, 2005). The intuitive system includes the ventromedial prefrontal cortex (vmPFC), medial temporal lobe (MTL) memory-processing structures, and distributed temporal lobe structures for supporting conceptual coherence and the implementation of language rules. The reflective reasoning system includes bilateral parietal cortex.

Whereas other theories of reasoning predict counterfactual inference and manipulation of mental representations as a major component, the *principle of truth* in MM theory – that it is easier to represent what exists by using mental models than what does not exist in fully explicit models – suggests that people will only engage in counterfactual reasoning when necessary, instead giving a central role to only the maintenance component of working memory in support of tracking multiple possibilities. This process engages the vmPFC, and could explain why people intuitively preferred calling a double-prevention relation a transitive prevention in some previous experiments (Barbey et al., 2011). This would be consistent with the dual-process hypothesis that causal reasoning includes both intuitive judgments (with simple mental models

that are consistent with prior beliefs) relying on the vmPFC and the combination of mental models to draw more complex conclusions using the parietal lobes for visuospatial manipulations (Goel, 2005). Such a pattern has been observed in studies focusing on deductive reasoning. Seeing a similar pattern in reasoning about causal relations that are not explicitly deductive would be consistent with deductive reasoning and mental models being descriptively valid in the context of complex causality, but would not rule out the possibility that deductive reasoning is primarily causal instead of causal reasoning being fundamentally deductive.

Recent developments in MM theory go beyond visuospatial manipulation, however, suggesting that the core features of the theory can also be mapped onto other brain structures (Khemlani et al., 2014). Specifically, the reflective system in the dual-processing framework mentioned above should be expanded beyond the parietal lobes to include the prediction that reasoning over mental models will engage the lateral prefrontal cortex (lPFC). Recall that mental models are modal in nature; they license possible states of affairs. The encoding of stimulus-response rules has been mapped to populations of neurons in dorsolateral PFC (dlPFC) (Mian et al., 2014); these stimulus-response mappings could be interpreted as an action-based instantiation of a more general mechanism in the lateral PFC for supporting mental models that link events in space and time. Although mental models are not iconic in perceptual form, they have been called iconic in that they preserve the relative structures of objects in space, events in time, and members of abstract sets to each other (Johnson-Laird et al., 2004). Subdivisions of lateral PFC are recognized as signaling object and concept representations that could plausibly support the maintenance of abstract sets of objects and events in the mind. Finally, the *principle of truth* suggests a natural preference for representing true states of affairs, rather than the alternative possibilities that can be imagined from a fully explicit set of models. The ability of

prefrontal neurons to maintain stimulus-related activity in the absence of a stimulus has long been a central principle of understanding the prefrontal cortex (Fuster, 1989). This feature is also the basis of the Guided Activation model of the PFC, a process model in which the PFC primarily serves a control function, selectively activating or inhibiting different stimulus-response mappings according to the contexts that dictate when they are appropriate or not. The fact that the PFC (lateral PFC in particular) must assign task-relevance to some representations over others and sustain attention to them in the service of goal-directed behavior, has been suggested as also supporting belief-oriented processes like reasoning (Khemlani et al., 2014). Under this view, sustaining attention to a task-relevant set of representation and discarding distractors is analogous to sustaining attention to a model of what is true while disregarding alternative possibilities.

Note that a role being played by lateral PFC in reasoning is not unique to MM theory. The neural correlates of cognitive flexibility feature prominently in this hypothesis, but alternative hypotheses emphasizing causal models or force representations would presumably appeal to cognitive flexibility as well (Barbey et al., 2013). What is specific to MM theory among the other models discussed here is its suggestion that a combination of abstract models and symbolic manipulations (e.g. truth statements and negations) is part of causal reasoning. Lateral PFC supports abstract symbolic manipulations of information being held in working memory, including information that is not primarily a sensory mapping or sensory reconstruction of the world (Khemlani et al., 2014; Ramnani & Owen, 2004; Tettamanti et al., 2008).

Together, the neural implications of MM theory can be summarized as the engagement of a reflective reasoning system including the right lateral prefrontal cortex and right parietal

cortex, with the interaction between prior beliefs and evidence supported by bilateral ventromedial prefrontal cortex and fronto-temporal memory systems.

3b. Causal Models Theory

One advantage of CM theory is that its implementation would place less of a demand on the number of slots available to working memory processes (or bits of information that can be represented). The combinatorial explosion characteristic of purely statistical accounts of causal judgment and reasoning is not entirely escaped by MM theory, when multiple relations require the expansion of mental models into their fully explicit models. When people reason correctly about the combination of two prevent relations into a single allow or cause relation, a graphical Bayes Net would place less of a demand on the limits of short-term memory, while also supporting the manipulation of variables' values and edges' directions to recognize the lack of transitivity. This process of intervention, central to the CM approach, requires the manipulation of information in working memory, and selectively relies on the dlPFC (Barbey, Koenigs, et al., 2012).

Furthermore, CM theory flexibly supports the representation of either a probabilistic or deterministic world-view, such that probabilistic data about a relation can be interpreted just as easily in terms of alternative causal nodes as in terms of a fundamentally stochastic relation. The neural correlates of inductive and deductive reasoning, however, remain to be well characterized (see Section 3a). As mentioned previously, some studies find that induction is specific to the left PFC with deduction being localized to the right hemisphere (Parsons & Osherson, 2001), whereas others find that induction and deduction are both left-lateralized, with ventral selectivity for deduction and dorsal selectivity for induction (Goel & Dolan, 2004). Other neuroimaging

studies from the decision-making literature, rather than those directly bearing on causal reasoning, have identified an uncertainty monitoring network engaging the PFC, parietal and insular lobes (Huettel et al., 2005). The neural correlates even appear to change according to the type of probability being represented; ventromedial PFC, insula, amygdala and putamen are increasingly activated when judging on the basis of uncertain prior probabilities, and activation in posterior occipital cortex scales up according to increasingly uncertain conditional probabilities (Vilares et al., 2012).

A key feature of CM theory is that it supports intervention – that is, causal models can be manipulated to reflect some alternative, or "counterfactual", state of affairs. This is known as counterfactual reasoning. Neuroscience theories of counterfactual reasoning suggest that the medial PFC plays a key role in imagining alternative states of affairs, with different regions supporting different types of manipulation (Barbey et al., 2009). The ventral medial PFC, often associated with value assignment and motivational representations, supports counterfactuals that differ in valence of value (better or worse than reality). Dorsal medial PFC supports the distinction between action and inaction. According to this view, dorsal medial PFC supports a general mechanism for representing states of affairs that do not currently match reality. This view is consistent with other accounts of dorsal medial prefrontal cortex (including the cingulate gyrus, a fold in the frontal cortex beneath the outer-most layer), that assign to it a central role of representing prediction error or conflict between expectancies and reality (Botvinick et al., 2001). Counterfactual reasoning will also be supported by the working memory mechanisms described above, by allowing multiple alternative states of affairs to co-exist while a particular state is being modeled and manipulated. The concurrent maintenance of multiple goals or action outcomes has been mapped to the frontopolar cortex, another name for the most anterior region

of the prefrontal cortex, with the two hemispheres dividing the labor in a task-switching paradigm (Charron & Koechlin, 2010).

On the basis of cognitive neuroscience theory concerning the functions that comprise reasoning through the use of Bayes Nets, the neural implementation of reasoning according to CM theory would engage neural correlates of counterfactual reasoning, working memory manipulation (rather than pure maintenance), probability judgments and explanatory reasoning to resolve uncertainty. We would thus expect to see a primarily left-hemispheric fronto-temporal network supporting causal reasoning.

3c. Force Composition Theory

The force representations in Force Composition Theory are based on iconic perceptual codes in that their organization reflects the structure of the relations being represented (Wolff & Barbey, 2015). By analogy, a subway tunnel map is an iconic representation of the true physical structure that preserves topology (while sacrificing topographical accuracy). If the iconic character of perceptual codes in causal reasoning is limited to their organization in relation to one another as represented in visual free-body diagrams, then the process of reasoning from causal premises to a conclusion would primarily engage neural mechanisms for constructing and manipulating symbolic visuospatial models, rather than isomorphic representations that reflect the details of actual causal relations. The superior parietal lobe serves as the node within a larger fronto-parietal working memory network that supports the manipulation of visuospatial models (Koenigs et al., 2009).

On the other hand, if the iconic nature of force representations goes as far as imagining visual re-creations of the way forces interact with one another in real life, then we would expect

to see involvement of more ventral, modality-specific neural correlates in the construction of simulations that are then "watched" in the mind's eye to predict how an imagined causal system would behave (Patterson & Barbey, 2012). Specifically, occipital and parietal cortex support the construction of visual simulations; premotor cortex, temporal cortex, and occipital cortex support representations of action and biological motion (Patterson & Barbey, 2012; Paus, 2005; Schacter et al., 2007). Biological motion may be particularly important to the construction of causal force representations involving agency, because the features discriminating biological motion as detected by more recently-evolved anterior structures in the brain from more primitive motion detectors in early visual processing are not purely structural. The movement of articulated joints and facial features are processed differently than pure motion because they signify the coherent, animated activity of an organism toward some goal or away from some consequence. This could be considered a very rudimentary form of causal judgment, and it could conceivably support causal reasoning through the use of mental simulations. If so, we would expect to see the neural correlates of agency and intention featuring prominently in causal reasoning neuroscience experiments: predominantly the middle temporal and medial superior temporal regions (typically designated by landmarks at the posterior end of the superior and inferior temporal sulci) (Grossman et al., 2000).

In summary, the neural correlates of force composition would primarily involve modality-specific engagement of sensory processing networks in occipital, parietal and posterior temporal cortex to support the creation of an iconic mental simulation, with right-sided parietal lobe engagement to support the manipulation or "running" of the simulation.

4. Review of the Cognitive Neuroscience Evidence on Causal Reasoning

This section will review the results of cognitive neuroscience research using experiments that involve thinking about causal relations (see Figure and Table 4 for a summary of the main findings from the fMRI studies on causal judgment and reasoning discussed here). The vast majority of cognitive neuroscience research on causal relations is focused on causal judgment, or inductively concluding that a causal relation exists. Here we distinguish causal judgment from the type of causal reasoning that the psychological models describe in greater depth: the combination of previously induced causal relations to infer a larger causal relation that has not been directly observed. Although judgment and reasoning can be thought of separately in this manner, the major claims of each theory can be applied to causal judgment as well. Causal judgment is a form of causal reasoning, in that it involves transforming one form of knowledge (the perception of events occurring together) into another (that the events are linked by some generative mechanism). Further neuroimaging research should focus on causal reasoning over multiple complex relations, such that the behavioral models of causal reasoning can be directly mapped onto neuroscience models of network activity, rather than indirectly inferred as in this chapter.

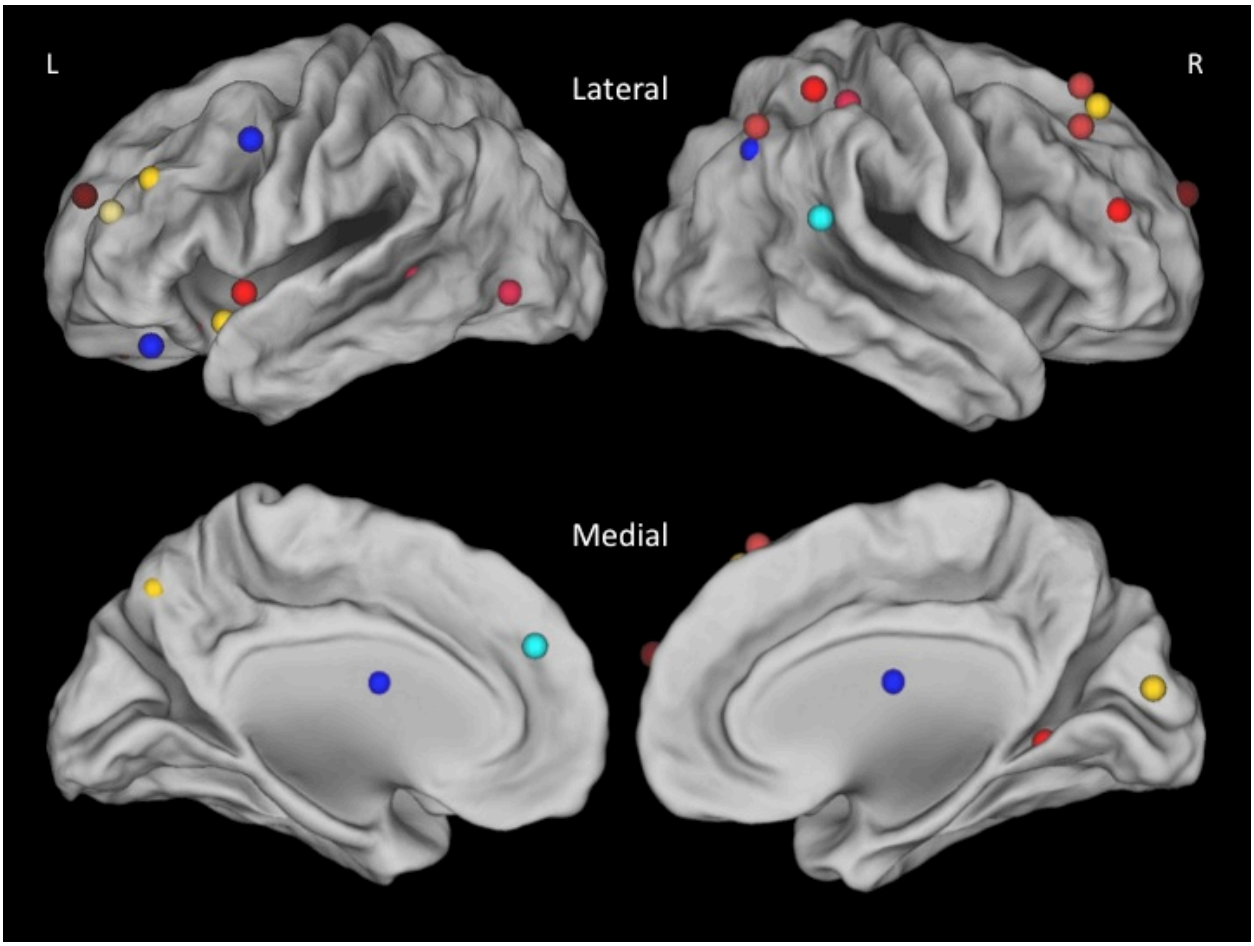


Figure 4. Brain activation foci from fMRI studies on causal judgment and reasoning. The colored spheres each represent the peak voxel in an activation cluster resulting from a linear contrast or regression model. Each color (or shade of color) represents a different study. Spheres in shades of red represent studies using Michotte-style collision stimuli. Spheres in shades of blue represent studies using social or interpersonal causal attribution stimuli. Spheres in shades of yellow represent studies using abstract or verbal causal task stimuli. Individual clusters with fewer than 10 voxels are excluded from this figure, as are cerebellar clusters and most subcortical clusters.

Study Name	Stimuli	Contrasts of Interest
Fonlupt, 2003 (dark red)	Physical collisions	• Causality > movement
Fugelsang et al. 2005 (light red)	Physical collisions	• Causality > causal violations
Straube and Chatterjee, 2010 (pink)	Physical collisions	• Activation increasing with sensitivity to violations of causality
Woods et al., 2014 (bright red)	Physical collisions	• Causal judgment (causal and non-causal events) > resting baseline
Fugelsang and Dunbar, 2005 (bright yellow)	Simulated medical treatment data	• Plausible > implausible medical explanation • Data inconsistent with plausible explanation > data consistent with plausible explanation • Data consistent with plausible

		explanation > data inconsistent with plausible explanation
Satpute et al., 2005 (light yellow)	Verbal pairs	• Causal relationship > non-causal association
Blackwood et al., 2003 (dark blue)	Vignettes	• Internal > external attribution • Self-serving bias > self-critical bias • Self-critical bias > self-serving bias
Harris et al., 2005 (light blue)	Vignettes	• Individual attribution > general attribution • Social attribution > general attribution

Table 4. fMRI studies on causal judgment and reasoning.

4a. Causal Judgments Concerning Physical Events

To study the neural mechanisms in the brain supporting the perception of causal relations among physical objects interacting with one another, study participants are asked to discriminate between causal and non-causal event chains in a series of videos or vignettes of objects colliding. Michotte's "launching events" are the original and most frequently used version of this testing paradigm, involving of billiard balls colliding with one another (Michotte, 1963; White, this volume).

One of the earliest neuroscience studies of causal judgment involved comparing two conditions: the first in which people were instructed to judge whether an event was causal or not, and a second in which they were instructed to judge which direction a particular ball moved (Fonlupt, 2003). Within each condition were two event types: a causal event in which one ball strikes another to launch it, and a non-causal event in which one ball simply passes a stationary ball without contacting it. The reason for this 2x2 study design was to settle a controversial question concerning the nature of causal judgments: is there an automatic "cause perceiving" module in the visual processing regions of the brain, similar to feature detectors or ensembles of neurons that respond to specific shapes in specific orientations? Or is causality something that

must be inferred by putting together the components of an image or vignette? By showing people videos of causal and non-causal events while asking them to alternate being attending to causal status or lower-level physical features, it is possible to test whether the "neural signature" of causal perception is activated any time a causal event is viewed regardless of attention and intent, or is only activated when trying to make a causal judgment. This study found no difference in brain activation between viewing causal and non-causal events within a particular block of the experimenters' instructions, but the process of trying to making a causal judgment, regardless of the stimulus's features, activated the medial prefrontal cortex when compared with the lower-level perceptual process of judging motion. There are a number of ways to interpret this finding, but the most commonly accepted one is that causal judgment is not an automatic result of low-level perceptual features activating a causality-specific module; even simple causal judgments are the result of a PFC-mediated conscious process in the brain.

Another study using Michotte's launching events focused only on the process of making causal judgments about events in which a moving ball approaches another ball before stopping and launching a stationary second ball (Fugelsang et al., 2005). The goal of this study was to investigate the neural mechanisms that enable people to use features like spatial and temporal contiguity between events to infer causality. The authors manipulated the stimuli such that the non-causal events still involved collisions, but included some violation of the normal rules of physics. The spatial gap condition involved the first ball coming to rest before making contact with the other ball, and the other ball beginning to move without having been touched by the first; if this happened in real life, we would assume some force other than the collision accounted for the second ball moving, even if somehow related to the first ball's movement (e.g. electric repulsion between subatomic particles with the same charge). The temporal gap condition

involved the first ball colliding with the second, but the second ball only beginning to move after a delay of several seconds. When comparing the neural activations of judging causal events with judging events featuring causal violation, the authors found neural activation in the right middle frontal gyrus (in the prefrontal lobe) and right inferior parietal lobule (in the parietal lobe). Comparing the causal condition to only the temporal delay condition selectively activated the right inferior parietal lobule. Comparing the causal condition to only the spatial gap condition activated the right middle temporal gyrus. Together, these results indicate that there is a predominantly right-hemisphere network for perceiving causality from physical events, with the parietal lobe being particularly sensitive to detecting spatial contiguity (or inactivation by temporal discontiguity) and the temporal lobe being sensitive to detecting temporal contiguity between events (or inactivation by spatial discontiguity).

The intriguing nature of the causal violation studies motivated another neuroimaging experiment using launching events. This study used materials in which causal violations were not in discrete categories unto their own, but instead involved gradually introducing violations of causality such that they were barely noticeable at first and only slowly became more extreme in two domains (Straube & Chatterjee, 2010). The temporal delay domain involved increasing the increment of time between the collision and the launching of a second ball, starting at zero. The spatial domain involved increasing the incident angle of the second ball's trajectory, starting at zero degrees from horizontal, until the balls began moving away at 90 degrees from the direction in which they were hit. Participants were asked to judge whether the collisions and launchings were causally linked while having the BOLD response in their brains imaged in an MRI scanner. Treating the conditions as categorical in the first analysis (causal vs. non-causal) resulted in no difference in neural activation between judging causal and non-causal stimuli. Instead, a general

causal judgment network was engaged by either condition when comparing the neural activity against a resting baseline with no visual stimulation; it included large areas of the occipital, parietal and frontal lobes. This is consistent with the earlier findings on causal judgment and perception of motion. The absence of a clear distinction between conditions could have been an artifact of there not being a clear boundary between causal and non-causal events in the study stimuli, however. Crucially, not all participants responded equally to the violations of causality. People who were more sensitive to temporal delays had greater activity in the left putamen, a subcortical structure associated with controlling movement and the construct of implicit memory, including motor and procedural memory. Individuals who were more sensitive to spatial violations showed greater activation in the right post-central gyrus and the right parietal lobule. The post-central gyrus is also known as the location of the sensory homunculus, allowing localization of touch, pain, proprioception and other aspects of bodily states. The parietal lobule is known to support spatial mapping and manipulation of objects in space. These findings suggest the emergence of a domain-general causal network supporting perceptual causal reasoning about physical events, with some specific nodes selectively active in the processing of particular features of causal relations, namely spatial and temporal violations of expectation.

A subsequent study used the same stimuli: launching events with increasingly extreme violations based on temporal delays and angles of impact. The key manipulation in this study was that participants were alternately instructed to focus only on one domain or the other – spatial or temporal contiguity – while making causal judgments (Woods et al., 2014). As with the previous neuroimaging studies, no difference was seen between the neural activations of responding to causal events and non-causal events. Comparing the general mental state of causal judgment against a resting baseline revealed activation in a largely right-sided network involving

the right inferior and middle temporal gyrus, right lingual gyrus, right caudate, bilateral putamen, bilateral insula, right parietal cortex, middle frontal gyrus, and bilateral cerebellum. When participants were asked to focus on the spatial properties of the events, those who were more sensitive to violations had increased activity in the bilateral inferior frontal gyrus, bilateral inferior parietal cortex, and right superior parietal cortex. When asked to focus on time, the participants who were more sensitivity to violations had greater activation in the right hippocampus and the bilateral vermis of the cerebellum. A confirmatory follow-up study used transcranial direct-current brain stimulation (tDCS) to test whether the responses to causal violations can be made more sensitive by selectively stimulating single brain regions from among those revealed in the fMRI analysis. tDCS is believed to have an effect by lowering the threshold for neuronal firing in the brain regions activated in the fMRI experiment, rather than actively causing neuronal firing (see (Nitsche et al., 2008) for a review of hypotheses concerning the effects of tDCS). Specifically, the authors used anodal stimulation of the right hemisphere, comparing three conditions: frontal lobe stimulation, parietal stimulation, and sham stimulation as a control condition. Frontal stimulation increased sensitivity to violations of either spatial consistency or temporal contiguity. Parietal stimulation increased sensitivity to violations of spatial contiguity only. Together, the fMRI and tDCS findings provide evidence for a right-sided network suggesting frontal support for general perceptual causality, and parietal sensitivity to the spatial properties of events and relations between them.

4b. Complex Judgments of Abstract Relations

We can imagine any number of complex events that cannot be adequately explained or predicted in terms of visuospatial representations of physical collisions. Judgments of agency

and intentionality in a conversation between friends or enemies, for example, require more nuanced explanations involving current evidence (statements), prior knowledge (personality traits and the likelihood of their causing particular statements), and an understanding of the fact that tone and context influence meaning as much as the semantic content of a conversational exchange. By using vignettes with people interacting with one another, or descriptions of events that need to be explained or predicted in the context of prior experience, the neural correlates of complex causal judgment and reasoning can be explored in comparison with the neural framework for supporting physical causality.

As alluded to by our prior examples of reasoning about the relations between smoking and lung cancer, the field of medical diagnostics and treatment planning is an area rich with opportunity for studying how people understand, represent and manipulate complex causal networks. One fMRI study was designed to explore the neural correlates of interactions between evidence and prior beliefs, especially as it pertains to the plausibility of a causal mechanism having the effect being predicted or explained (Fugelsang & Dunbar, 2005). Specifically, it measured the neural correlates associated with judging the efficacy of two treatments for depression (either a plausible pharmacological treatment or placebo) in the context of two statistical patterns (low rate of treatment success, or high rate of treatment success). Participants were asked to decide how effective each of the treatments was in predicting happiness after having seen 20 individual trials of each condition (each trial being a hypothetical patient who was administered the drug and either responded or failed to respond to the treatment). The authors correctly predicted that causal attributions would be highest in the condition featuring a plausible mechanism (not placebo) and high treatment response rates. The use of correlation data

alone in causal judgment, as simulated by the placebo without a plausible mechanism, appears inadequate to infer a causal mechanism.

Activity in several brain regions was observed when comparing the BOLD response corresponding to the different study conditions. First, the authors compared the consideration of plausible theories with implausible theories, regardless of treatment response rates. The left inferior frontal gyrus, right superior frontal gyrus, and primary visual cortex were all activated when contrasting participants' consideration of plausible theories against their consideration of implausible ones. The authors suggest that this provides evidence for the involvement of working memory, executive control and visual attention mechanisms that have been previously attributed to these regions. Within each plausibility condition (medicine or placebo), different activations were also seen when directly contrasting the blocks with treatment response rates that are consistent or inconsistent with prior knowledge. Treatment response rates that are consistent with prior beliefs about the causal mechanism (treatment success after medicine) selectively engaged the left parahippocampal gyrus (PHG) and right caudate nucleus, whereas data conflicting with the plausible mechanism (taking medicine, but no response) selectively engaged the right cingulate gyrus, left dorsolateral prefrontal cortex (dlPFC) and left superior parietal lobe. Surprising treatment rates were generally disregarded in the implausible, or placebo, condition. The fact that medial temporal lobe structures (of which the PHG is one) are overwhelmingly implicated in the processing of episodic memory and semantic knowledge suggests that memory representations are retrieved when considering the plausibility of a theory; the error-monitoring and conflict-monitoring role often attributed to the cingulate cortex and dlPFC suggests that inconsistent data in the context of a plausible mechanism is manifested as a prediction error of sorts. Greater activation of the left hemisphere during conflicts between theory and evidence

provides support for the counterfactual or Causal Models approach to causal reasoning, and right-sided frontal activations while participants evaluated evidence consistent with an implausible theory could support an inhibitory or conflict-monitoring function under any of the theories of causal reasoning.

It is worth noting here that the multiplicity of functions supported by a given structure or network in the brain renders the reverse inference approach to understanding brain function (inferring which functions or calculations are being used to complete a task on the basis of seeing a particular neural pattern of activation) less than exhaustive (Poldrack, 2006). The fact that a particular brain region has been associated with some known function in the past does not necessitate that it must be fulfilling the same function in all subsequent tasks that engage it; many brain regions have the ability to support more than one function. This is not a criticism unique to the methods reported here, and is instead a limitation inherent to exploratory analyses in cognitive neuroscience. Still, the reverse inference approach is a reasonable starting point for hypothesis development in a field as young as the cognitive neuroscience of causal reasoning. This is especially the case when dealing with such reliably observed function-mappings as medial temporal lobe memory processing, and executive control in dlPFC.

4c. Social Causal Reasoning

Social and emotional intelligence are burgeoning areas of study in the cognitive and neural sciences (Hilton, this volume). An open question concerning the structure of social intelligence and its relation to other constructs of human brain function is whether it is a unique set of faculties specific to processing social and emotional information, or simply an aspect of general intelligence that emerges when the content of representations being supported happens to

feature social and emotional information. Identifying the neural basis of social cognition will not end the debate, but the fact that some studies reveal an independent set of competencies specialized for social cognition suggests that there should be unique neural contributions to such competencies.

Explaining and predicting the behaviors of others (and our own behaviors) are two processes requiring the representation and manipulation of causal relations involving stable attributes, intentional states and actions of people as they interact with one another. By asking people to make judgments about the actions of others while being imaged in an fMRI study, several studies have begun to separate the layers of social causal cognition into the types of information that are central to this type of reasoning beyond physical collisions or impersonal probabilistic events.

One study examined the types of information that influence how we generate causal explanations for human behavior. People are presented with vignettes, and asked to draw one of four conclusions: the behavior was due to the main actor's characteristics, the behavior was due to the characteristics of another person in the exchange, the behavior was due to impersonal contextual factors, or it was due to some combination of the three (Harris et al., 2005). According to the authors, people tend to make attribution judgments concerning an actor's behavior on the basis of information about consensus (whether other people act similarly), distinctiveness (whether the behavior is specific to a particular object, or all members of a target category), and consistency (whether the behavior is reliably seen by this actor). People are also most likely to attribute an action to a person's individual characteristics when consistency is high, but consensus and distinctiveness are low. For example, a kind gesture will be attributed to the personality of the actor if that person is routinely kind, especially in situations in which

others might not be, and if the person acts that way indiscriminately with respect to whom is receiving the kindness.

In this experiment, brain activity was measured using fMRI while participants engaged in the attribution task. Activity in the superior temporal cortex (STS) was elevated in the combination of conditions evoking person-attribution (low consensus and distinctiveness, high consistency), when compared with the other combinations of conditions. Activity in the mPFC, but not STS, was also associated with social judgments not specific to a single person (high consensus, low distinctiveness, high consistency). The person-attribution condition also activated other regions in the brain (right middle temporal gyrus, right middle occipital gyrus, right precentral gyrus, right precuneus, left insula and left cingulate gyrus), but not uniquely when compared with other study stimulus conditions (e.g. low consensus, distinctiveness and consistency). The fact that right STS and left mPFC are preferentially engaged by social cognition and Theory of Mind test paradigms suggests that social causal reasoning converges with the ability to infer the mental states of others. The left-sided prefrontal activation is consistent with CM theory, but could conceivably be an artifact of the social component of reasoning rather than causal attribution *per se*. The right-sided temporal activation is consistent with the biological motion (or agency) and spatial reasoning aspects of FC theory, but could also hypothetically be part of a larger right-sided activation pattern characteristic of deductive reasoning, as seen in MM theory. Note that the areas that are activated together by multiple study conditions are primarily right-hemispheric, which would provide support for MM theory.

Biased attribution of behavior on the basis of trying to serve some personal motivation also has a rich literature demonstrating exactly how error-prone and flexible human social judgment can be (Kunda, 1990). In Western cultures that assign a high value to individualistic

notions of self, we tend to erroneously attribute the actions of others to their dispositions while underweighting the influence of context; this was famously termed Fundamental Attribution Error, and it features prominently in the social psychology literature (Mason & Morris, 2010; Ross, 1977). Similarly, people often make overly-forgiving judgments of their own actions, taking credit for successes and blaming circumstance for failures and indiscretions, presumably to reduce dissonance and preserve a positive self-image (Greenberg et al., 1982).

As with the other tendencies to construct causal models as dictated by our goals and contextual factors, the presence of attribution biases can be mapped to a network in the brain supporting its component parts: in this case, general causal reasoning and mechanisms for representing the assignment of value to particular explanations (Blackwood et al., 2003). By instructing participants to imagine themselves as the central actor in a series of social vignettes requiring an explanation for their own behaviors, it was possible to directly compare the neural correlates of self-attribution and other-attribution. Participants could choose from self-attribution, other-attribution, or situational factors. Attributing actions to the self without a bias (so, including negative and positive actions) engaged the left lateral cerebellum, bilateral dorsal premotor cortex, and right lingual gyrus. External attribution (collapsing other person and impersonal contextual influences into a single category) engaged the left posterior superior temporal sulcus (STS). Comparing the activations associated with a self-serving bias with those associated with a self-deprecating bias revealed that favorable attributions activated the bilateral caudate nuclei, while a bias against self-serving attributions activated the left lateral OFC, right angular gyrus, and right middle temporal gyrus. The role of the STS in general external attribution is attributed to its role in inferring the mental or intentional states of others, and the role of premotor cortex, cerebellum and lingual gyrus in general self-attribution is linked to their

role in simulating one's own actions and intention states in decision-making. The neural basis of these biases is of particular interest, because the very presence of a bias suggests some error in reasoning – a departure from rational thought that might help explain what makes humans unique. Activation of the caudate nucleus when making self-serving attributions suggests that representations of reward and motivation are involved. It is conceivable that multiple causal representations are concurrently constructed in this context: first, a plausible causal model linking self and situational factors to the behavior in consideration, and secondly, an implicit causal model linking the very inference being drawn to a particularly desirable emotion state, accounting for why the self-serving bias is even observed in the first place. The engagement of bilateral frontal, temporal and parietal lobes in any internal attribution is consistent with all three theories of causal reasoning. Engagement of the lingual gyrus and parietal lobes in particular, especially in contrasts not involving a major difference in visual processing, supports an iconic sensory modality-specific representation as suggested by FC theory. Bilateral caudate activation associated with a self-serving bias and a left frontal, right parietal activation in the self-deprecating bias are not clearly supportive of any one theory of reasoning. The left-sided temporal engagement of external attribution when contrasted with all other attribution types appears consistent with CM theory, but only when considered on its own without the context of the other conditions.

Evidence from causal reasoning experiments enrolling participants with brain damage serves to confirm several general trends in the neuroscience of social causal attribution. Between two otherwise-equivalent explanations for an event of interest, healthy adults will tend to favor the explanation featuring agency or intention on the part of some involved person. Accumulating evidence suggests that some patterns of brain damage are more likely than others to impair the

discrimination between intentional and unintentional acts in their causal power (Channon et al., 2010). The study stimuli involved asking participants to read chains of events with two causes preceding some effect, with each cause varying from intentional or unintentional human acts to physical events not involving humans at all. Then, participants were asked to rate the causal power of each cause in the chain on a four-point likert scale, before using a similar scale to decide which cause was central to the effect. When comparing neurological patients with frontal damage (especially in the right hemisphere) to those with posterior damage and healthy control subjects, it appears that the frontally damaged group is still able to discern the two, but to a lesser extent than participants in the other groups. Specifically, right middle frontal gyrus, right inferior frontal gyrus, right ventrolateral PFC (vlPFC) and right insular cortex damage predicted a lesser extent of discrimination between intentional and unintentional human acts in their causal power. The findings provide evidence for an anterior right-hemisphere network that is critical to the discrimination between acts made intentionally or unintentionally. This pattern of lesion-symptom mapping is consistent with the MM view of deductive inference as the driving force behind causal attribution.

4d. Causal Judgment Versus Associative Learning

The final element of causal judgment that has been studied using neuroscience methods is the distinction between associative learning mechanisms and causal reasoning. In integrating cognitive and behavioral psychology with the rich philosophical tradition on models of reasoning, connections can be drawn between such constructs as goal-directed behavior and causal representations; goal-directed behavior would be incoherent without some understanding of causality to predict the consequences of actions and adapt behavior accordingly. One hurdle to

explaining brain function and higher cognition in terms of causal representations is that associative learning mechanisms along the lines of classical conditioning can explain animal behavior that resembles an understanding of causal relations.

Associative learning is based on tracking patterns of coincidence, and although it might engage semantic knowledge representations for the objects being associated, there should be no need to engage semantic knowledge to describe the nature of relations if causality truly exists as a privileged class of representation beyond associations (see LePelley, Griffiths, & Beesley, this volume, for associative theories of causality; Boddez, DeHouwer, & Beckers, this volume, for reasoning theory). To study the neural correlates engaged by causal judgment beyond the semantic knowledge network in the brain supporting associative judgment, one fMRI study instructed participants to judge the nature of a series of paired words while being scanned (Fenker et al., 2005; Satpute et al., 2005). The pairs of words varied as to whether they were causally linked (e.g. wind and erosion), non-causally associated (e.g. ring and emerald) or unrelated (e.g. eggs and liar). The study was divided into two block types that involved manipulating whether participants were instructed to judge each pair as causal versus unrelated, or associated versus unrelated. In the causal judgment condition, then, associated pairs would warrant a "no" response, while causal pairs in the associative condition would still warrant a "yes," because causal links are simply one type of association in this context. Note that semantic knowledge is still required to make an associative judgment, but not what the authors call a "role binding" process assigning each word in the pair to either the cause or effect role. The neural correlates of this role binding process emerged when common features of causal and associative judgment were subtracted from those associated with causal judgment only. A mostly left-hemisphere semantic processing network emerged when combining the two conditions: left

dorsolateral prefrontal cortex (dlPFC), left middle frontal gyrus, inferior frontal gyrus, superior parietal lobule, anterior cingulate gyrus, fusiform gyrus, and bilateral cerebellum. The neural correlates of causal reasoning contrasted against associative reasoning were much more focal, including a more anterior cluster in dlPFC and the right precuneus. Associative reasoning selectively engaged right superior temporal gyrus (STG) when contrasted with causal-only judgment.

5. Discussion

At first glance, it is tempting to divide the psychological theories of reasoning according to the dissociable brain networks that they would predict as those supporting causal reasoning. Mental Models Theory emphasizes deduction over an abstract code of possible states of affairs, which generally engages a left-hemisphere fronto-parietal network. Causal Models Theory emphasizes inductive reasoning over an abstract network representing both statistical dependencies and generative mechanisms linking the variables in a set of counterfactual manipulations, which should engage a right-hemisphere frontal-temporal-parietal network. Force Composition Theory emphasizes iconic force vector representation, which could take the form of symbolic free-body diagrams or life-like representational simulations. This would suggest a network involving the superior parietal lobe, early perceptual processing streams in parietal and occipital lobes, and medial prefrontal cortex.

On the basis of the causal judgment neuroscience literature alone, there appears to be broad evidence for a frontal-temporal-parietal network, in support of a plurality of cognitive or psychological models. Particularly in the context of necessary and sufficient relations that are considered transitive according to formal logic (e.g. cause and enable), joining multiple relations

occurs using a deductive process by definition, which would suggest engagement of the right hemisphere, and the use of mental models. Uncertainty and probability could plausibly engage a separate, inductive mechanism of reasoning using Bayes Nets in the left hemisphere, or it could still proceed using a fundamentally deductive mechanism as suggested by Johnson-Laird and colleagues (Johnson-Laird, 1994). The divided tracking of multiple outcomes in a decision making task is also known to require bilateral representation in the brain (Charron & Koechlin, 2010). Furthermore, reasoning about abstract causal relations like intention on the part of agents interacting with one another seem to support the face validity of a simulation approach to causal reasoning when modal and statistical network based representations do not quite capture the nature of the causal mechanism. Simplifying a series of abstract "forces" as a system to moves toward or away from an endstate is a plausible mechanism that may be engaged when necessary. In summary, there is most likely a plurality of modes of causal judgment and reasoning that are available. They may not all function at once, and may not even describe the same processes in the brain, but could rather be selectively recruited when a situation requires it. Reasoning about deterministic concepts – whether truly deterministic in nature or not – will likely involve deduction over mental models. Reasoning over probabilistic co-dependencies will likely involve induction over causal models. Reasoning about scenarios involving agency and complex instances of multiple preventions co-occurring will likely involve the use of iconic force representations.

6. Questions for Future Study

It is important to consider the questions that remain unanswered by the early research findings to date. In the carefully controlled environment of a psychology lab, there are a number

of features of causal reasoning that can be readily manipulated. Briefly, the difference between diagnostic and predictive reasoning has been the subject of some inquiry from a purely psychological standpoint (see Meder & Mayrhofer, this volume). Causal action simulations plausibly support predictive reasoning, for example. Diagnostic reasoning, however, appears to rely on similar information about causal dependencies and mechanisms, and it has been suggested that causal simulations cannot be run in reverse to generate explanations (Fernbach et al., 2011; Meder et al., 2014; Sloman & Lagnado, 2015); it remains undecided whether a series of alternate causal simulations can be run in the forward direction to decide on an explanatory inference (see Lombrozo, this volume, for a discussion on explanation), along the lines of the multiple models featured prominently in Mental Models Theory (Goldvarg & Johnson-Laird, 2001) or the structural equations in Causal Models Theory (Sloman & Lagnado, 2015). Mapping this distinction to key events or networks in the brain will help answer an old question in cognitive neuroscience: is the reasoning process best described by a single, domain-general neural mechanism that is distinct from lower level information processing steps, or instead by separate domain-specific neural mechanisms that include the neural correlates of the perceptual processing dictated by the content of the relevant information (e.g. emotion processing, or visual spatial information).

Another open question is to what extent other cognitive functions can be re-explained in terms of causality. Action representations supporting goal-directed behavior represent causal knowledge of the consequences of possible actions. The ability to construct coherent categories with meaningful boundaries relies on a basis of theoretical knowledge over and above any sort of feature combination or exemplar model; this "Theory Theory" is fundamentally causal in nature as well (Rehder, 2003; Rehder, this volume: a, b). More fundamental cognitive functions like

attention and memory may not be intrinsically causal, but they clearly support causal reasoning. The exact nature of this relationship in the brain remains to be seen.

7. Conclusion

We return to the broader, theory-relevant questions left unanswered by prior research. Definitively confirming or rejecting the neuroscience predictions of descriptive models of causal reasoning will require neuroimaging studies that use identical materials to those used previously in behavioral experiments. Let us not forget that all models are, by definition, incomplete and therefore inaccurate. They involve simplifying assumptions on some domains to allow perturbations on other domains of interest to be studied. None of the descriptive models of causal reasoning is likely to encapsulate all of human causal reasoning – even if a unified descriptive theory should be developed. For this reason, the short-term goal of neuroimaging studies on causal reasoning should be to use experimental materials that resonate with the personal experiences that most people have in trying to predict the future and explain the past. And once we have more thoroughly mapped the causal judgment and complex causal reasoning to networks in the brain, we can begin to tackle the more daunting task of uniting all brain function together in a single code. Some have proposed that there is a common thread of information processing or representation types that unites the different elements of intelligence, reasoning and perception in the human brain (Christoff & Gabrieli, 2000; Duncan, 2001, 2010; Koechlin et al., 2003; Miller & Cohen, 2001; O'Reilly, 2010). We suggest that the common thread is the representation of causal relations, and eagerly await the next findings to confirm our hypothesis or situate it within a more all-encompassing framework.

References

Ahn, W., & Kalish, C. W. (2000). The Role of Mechanism Beliefs in Causal Reasoning. In *Explanation and Cognition* (pp. 199–225).

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, 299–352.

Alivisatos, B., & Petrides, M. (1997). Functional activation of the human brain during mental rotation. *Neuropsychologia*, *35*, 111–118.

Banich, M. T. (2009). Executive Function: The Search for an Integrated Account. *Current Directions in Psychological Science*, *18*(2), 89–94.

Barbey, A. K., Colom, R., & Grafman, J. (2013). Architecture of cognitive flexibility revealed by lesion mapping. *NeuroImage*, *82*, 547–554. doi:10.1016/j.neuroimage.2013.05.087

Barbey, A. K., Colom, R., Solomon, J., Krueger, F., Forbes, C., & Grafman, J. H. (2012). An integrative architecture for general intelligence and executive function revealed by lesion mapping. *Brain*, *135*(Pt 4), 1154–64. doi:10.1093/brain/aws021

Barbey, A. K., Koenigs, M., & Grafman, J. H. (2011). Orbitofrontal contributions to human working memory. *Cerebral Cortex (New York, N.Y. : 1991)*, *21*(4), 789–95. doi:10.1093/cercor/bhq153

Barbey, A. K., Koenigs, M., & Grafman, J. H. (2012). Dorsolateral prefrontal contributions to human working memory. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, 1–11. doi:10.1016/j.cortex.2012.05.022

Barbey, A. K., Krueger, F., & Grafman, J. H. (2009). Structured event complexes in the medial prefrontal cortex support counterfactual representations for future planning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1521), 1291–300. doi:10.1098/rstb.2008.0315

Barbey, A., & Wolff, P. (2002). Causal Reasoning from Forces. In L. Erlbaum (Ed.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (p. 2439). Mahwah, NJ.

Blackwood, N. J., Bentall, R. P., Ffytche, D. H., Simmons, a, Murray, R. M., & Howard, R. J. (2003). Self-responsibility and the self-serving bias: an fMRI investigation of causal attributions. *NeuroImage*, *20*(2), 1076–85. doi:10.1016/S1053-8119(03)00331-8

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict Monitoring and Cognitive Control. *Psychological Review*, *108*(3), 624–652. doi:10.1037//0033-295X.I08.3.624

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, *1124*, 1–38. doi:10.1196/annals.1440.011

Channon, S., Lagnado, D., Drury, H., Matheson, E., Fitzpatrick, S., Shieff, C., et al. (2010). Causal Reasoning and Intentionality Judgments After Frontal Brain Lesions. *Social Cognition*, *28*(4), 509–522. doi:10.1521/soco.2010.28.4.509

Charron, S., & Koechlin, E. (2010). Divided representation of concurrent goals in the human frontal lobes. *Science (New York, N.Y.)*, *328*(5976), 360–3. doi:10.1126/science.1183614

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405. doi:10.1037//0033-295X.104.2.367

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*(4), 545.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*(1-2), 83–120. doi:10.1016/0010-0277(91)90047-8
Christoff, K., & Gabrieli, J. D. E. (2000). The frontopolar cortex and human cognition: Evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. *Psychobiology*, *28*(2), 168–186.
Corballis, P. M. (2003). Visuospatial processing and the right-hemisphere interpreter. *Brain and Cognition*, *53*, 171–176. doi:10.1016/S0278-2626(03)00103-9
Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, *2*(November), 820–829.
Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–9. doi:10.1016/j.tics.2010.01.004
Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, *23*, 475–83. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11006464
Fenker, D. B., Waldmann, M. R., & Holyoak, K. J. (2005). Accessing causal relations in semantic memory. *Memory & Cognition*, *33*(6), 1036–1046. doi:10.3758/BF03193211
Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology*, *140*(2), 168–185.
Fonlupt, P. (2003). Perception and judgement of physical causality involve different brain structures. *Cognitive Brain Research*, *17*, 248–254.
Fugelsang, J. A., & Dunbar, K. N. (2005). Brain-based mechanisms underlying complex causal thinking. *Neuropsychologia*, *43*(8), 1204–13. doi:10.1016/j.neuropsychologia.2004.10.012
Fugelsang, J. A., Roser, M. E., Corballis, P. M., Gazzaniga, M. S., & Dunbar, K. N. (2005). Brain mechanisms underlying perceptual causality. *Brain Research. Cognitive Brain Research*, *24*(1), 41–7. doi:10.1016/j.cogbrainres.2004.12.001
Fugelsang, J. A., & Thompson, V. A. (2003). A dual-process model of belief and evidence interactions in causal reasoning. *Memory & Cognition*, *31*(5), 800–815. doi:10.3758/BF03196118
Fuster, J. (1989). *The Prefrontal Cortex*. New York, NY: Raven.
Goel, V. (2005). Cognitive Neuroscience of Deductive Reasoning. In K. Holyoak & R. Morrison (Eds.), *Cambridge Handbook of Thinking & Reasoning*.
Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Sciences*, *11*(10), 435–41. doi:10.1016/j.tics.2007.09.003
Goel, V., Buchel, C., Frith, C., & Dolan, R. J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *NeuroImage*, *12*, 504–514. doi:10.1006/nimg.2000.0636
Goel, V., & Dolan, R. J. (2004). Differential involvement of left prefrontal cortex in inductive and deductive reasoning. *Cognition*, *93*(3), B109–21. doi:10.1016/j.cognition.2004.03.001
Goel, V., Gold, B., Kapur, S., & Houle, S. (1998). Neuroanatomical correlates of human reasoning. *Journal of Cognitive Neuroscience*, *10*, 293–302. doi:10.1162/089892998562744
Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: a mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*(4), 565–610. doi:10.1207/s15516709cog2504_3
Greenberg, J., Pyszczynski, T., & Solomon, S. (1982). The Self-Serving Attributional Bias : Beyond Self-Presentation. *Journal of Experimental Social Psychology*, *18*(56-67).
Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., et al. (2000).

Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience*, *12*, 711–720. doi:10.1162/089892900562417
Harris, L. T., Todorov, A., & Fiske, S. T. (2005). Attributions on the brain: neuro-imaging dispositional inferences, beyond theory of mind. *NeuroImage*, *28*(4), 763–9. doi:10.1016/j.neuroimage.2005.05.021
Huettel, S. A., Song, A. W., & McCarthy, G. (2005). Decisions under uncertainty: probabilistic context influences activation of prefrontal and parietal cortices. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *25*(13), 3304–11. doi:10.1523/JNEUROSCI.5070-04.2005
Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition*, *50*, 189–209. doi:10.1016/0010-0277(94)90028-0
Johnson-Laird, P. N. (1995). Mental Models, Deductive Reasoning, and the Brain. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (pp. 999–1008). Cambridge, MA: MIT Press.
Johnson-Laird, P. N. (2010a). Against logical form. *Psychologica Belgica*, *50*(3 & 4), 193–221.
Johnson-Laird, P. N. (2010b). Mental models and human reasoning. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 18243–18250. doi:10.1073/pnas.1012933107
Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, *111*(3), 640–61. doi:10.1037/0033-295X.111.3.640
Jordan, K., Heinze, H. J., Lutz, K., Kanowski, M., & Jäncke, L. (2001). Cortical activations during the mental rotation of different visual objects. *NeuroImage*, *13*, 143–152. doi:10.1006/nimg.2000.0677
Khemlani, S. S., Barbey, A. K., & Johnson-laird, P. N. (2014). Causal reasoning with mental models, *8*(October), 1–15. doi:10.3389/fnhum.2014.00849
Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science (New York, N.Y.)*, *302*(5648), 1181–5. doi:10.1126/science.1088545
Koenigs, M., Barbey, A. K., Postle, B. R., & Grafman, J. H. (2009). Superior parietal cortex is critical for the manipulation of information in working memory. *The Journal of Neuroscience*, *29*(47), 14980–6. doi:10.1523/JNEUROSCI.3706-09.2009
Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology. General*, *136*(3), 430–50. doi:10.1037/0096-3445.136.3.430
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.
Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W.H. Freeman and Company.
Mason, M. F., & Morris, M. W. (2010). Culture, attribution and automaticity: A social cognitive neuroscience view. *Social Cognitive and Affective Neuroscience*, *5*(2-3), 292–306. doi:10.1093/scan/nsq034
Meder, B., Mayrhofer, R., & Waldmann, M. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*(3), 277–301.
Mian, M. K., Sheth, S. a, Patel, S. R., Spiliopoulos, K., Eskandar, E. N., & Williams, Z. M. (2014). Encoding of rules by neurons in the human dorsolateral prefrontal cortex. *Cerebral Cortex (New York, N.Y. : 1991)*, *24*(3), 807–16. doi:10.1093/cercor/bhs361
Michotte, A. (1963). *The Perception of Causality*. London, England: Methuen.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202. doi:10.1146/annurev.neuro.24.1.167

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, a H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. doi:10.1006/cogp.1999.0734

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316.

Nitsche, M. A., Cohen, L. G., Wassermann, E. M., Priori, A., Lang, N., Antal, A., et al. (2008). Transcranial direct current stimulation: State of the art 2008. *Brain Stimulation*, *1*(3), 206–23. doi:10.1016/j.brs.2008.06.004

O'Reilly, R. C. (2010). The What and How of prefrontal cortical organization. *Trends in Neurosciences*, *33*(8), 355–61. doi:10.1016/j.tins.2010.05.002

Osherson, D., Perani, D., Cappa, S., Schnur, T., Grassi, F., & Fazio, F. (1998). Distinct brain loci in deductive versus probabilistic reasoning. *Neuropsychologia*, *36*(4), 369–376. doi:10.1016/S0028-3932(97)00099-7

Parsons, L. M., & Osherson, D. (2001). New Evidence for Distinct Right and Left Brain Systems for Deductive versus Probabilistic Reasoning. *Cerebral Cortex*, *11*(10), 954–65. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11549618

Patterson, R., & Barbey, A. K. (2012). A Cognitive Neuroscience Framework for Causal Reasoning. In J. H. Grafman & F. Krueger (Eds.), *The Neural Representation of Belief Systems* (pp. 76–120). New York, NY: Psychology Press.

Paus, T. (2005). Mapping brain maturation and cognitive development during adolescence. *Trends in Cognitive Sciences*, *9*(2), 60–8. doi:10.1016/j.tics.2004.12.008

Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge, MA: MIT Press.

Poeppel, D., & Embick, D. (2004). Defining the relation between linguistics and neuroscience. *Linguistics*, (1), 1–16. doi:citeulike-article-id:6138571

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63. doi:10.1016/j.tics.2005.12.004

Ramnani, N., & Owen, A. M. (2004). Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. *Nature Reviews. Neuroscience*, *5*(March), 184–194. doi:10.1038/nrn1343

Ratcliff, G. (1979). Spatial thought, mental rotation and the right cerebral hemisphere. *Neuropsychologia*, *17*, 49–54. doi:10.1016/0028-3932(79)90021-6

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29*(6), 1141–59. doi:10.1037/0278-7393.29.6.1141

Ross, L. (1977). The intuitive psychologist and his shortcomings: distortions in the attribution process. *Advances in Experimental Social Psychology*, *10*, 173–220.

Satpute, A. B., Fenker, D. B., Waldmann, M. R., Tabibnia, G., Holyoak, K. J., & Lieberman, M. D. (2005). An fMRI study of causal judgments. *The European Journal of Neuroscience*, *22*(5), 1233–8. doi:10.1111/j.1460-9568.2005.04292.x

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: the prospective brain. *Nature Reviews. Neuroscience*, *8*, 657–661. doi:10.1080/08995600802554748

Sloman, S. A., Barbey, A. K., & Hotaling, J. M. (2009). A causal model theory of the meaning of

cause, enable, and prevent. *Cognitive Science*, *33*(1), 21–50. doi:10.1111/j.1551-6709.2008.01002.x

Sloman, S., & Lagnado, D. (2015). Causality in Thought. *Annual Review of Psychology*, *66*, 3.1–3.25.

Smith, D. V, Clithero, J. A., Rorden, C., & Karnath, H.-O. (2013). Decoding the anatomical network of spatial attention. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(4), 1518–23. doi:10.1073/pnas.1210126110

Straube, B., & Chatterjee, A. (2010). Space and time in perceptual causality. *Frontiers in Human Neuroscience*, *4*(April), 28. doi:10.3389/fnhum.2010.00028

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. *Advances in Neural Information Processing Systems*, 43–50.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science (New York, N.Y.)*, *331*, 1279–1285. doi:10.1126/science.1192788

Tettamanti, M., Manenti, R., Della Rosa, P. a., Falini, A., Perani, D., Cappa, S. F., et al. (2008). Negation in the brain: Modulating action representations. *NeuroImage*, *43*, 358–367. doi:10.1016/j.neuroimage.2008.08.004

Ungerleider, L. G., & Mishkin, M. (1982). Two Cortical Visual Systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of Visual Behavior* (pp. 549–586). Cambridge, MA; London, England: MIT Press.

Vilares, I., Howard, J. D., Fernandes, H. L., Gottfried, J. A., & Kording, K. P. (2012). Differential representations of prior and likelihood uncertainty in the human brain. *Current Biology*, *22*(18), 1641–1648. doi:10.1016/j.cub.2012.07.010.Differential

Vincent, J. L., Kahn, I., Snyder, A. Z., Raichle, M. E., & Buckner, R. L. (2008). Evidence for a frontoparietal control system revealed by intrinsic functional connectivity. *Journal of Neurophysiology*, *100*(6), 3328–42. doi:10.1152/jn.90355.2008

Wolff, P., & Barbey, A. K. (2015). Causal reasoning with forces - authors' proof. *Frontiers in Human Neuroscience*, *9*.

Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For Want of a Nail: How Absences Cause Events. *Journal of Experimental Psychology: General*, *139*(2), 191–221. doi:10.1037/a0018129

Woods, A. J., Hamilton, R. H., Kranjec, A., Minhaus, P., Bikson, M., Yu, J., et al. (2014). Space, time, and causality in the human brain. *NeuroImage*, *92*, 285–297.